Retributivism after Anselm

Alex Tuckness and John M. Parrish
Iowa State University and Loyola Marymount University

This article describes the relationship between the distinctive account of a retributivist theory of divine punishment developed in St. Anselm's doctrine of the atonement, on the one hand, and the development of later retributivist theories of temporal punishment that emerged in succeeding centuries among modern moral and political philosophers, on the other. We claim that this development took a far more circuitous and indirect path than has generally been appreciated by either theorists of punishment or historians of ethics and political thought. Indeed we suggest that the contemporary notion of strict retributivism as applied to temporal affairs emerged no earlier than the eighteenth century in the writings of Immanuel Kant. Drawing on key contributions by his immediate predecessors Samuel Clarke, Richard Price, and Adam Smith, we argue that Kant adapted Anselm's theology to reach political conclusions about punishment that his medieval and early modern successors had largely rejected: that in temporal as well as divine punishment, punishment must satisfy the standard of strict retributivism.

On its face, the title of this article, "Retributivism after Anselm," might seem foolhardy in its breadth. As commonly understood in contemporary philosophical discourse, "retributivism" embraces the full range of views that understand punishment as intrinsically rather than instrumentally valuable, a fitting and necessary response to wrongdoing. One distinctive feature of this approach is that mercy is characterized as treating people better than they deserve while retributive justice requires treating them precisely as deserve. How could the whole history of retributivism after Anselm be addressed in a single essay? We have argued in our book *The Decline of Mercy in Public Life*, that at least as applied to state punishment—that is, temporal, human, public punishment—retributivism was not the primary way of thinking about governmental punishment in western political discourse prior to the eighteenth century.[1] Immanuel Kant, drawing on a handful of predecessors such as Samuel Clarke, Richard Price, and Adam Smith, and more importantly, key tenets from Anselm's philosophy that had up to that point been associated almost exclusively with *divine* eternal punishment, more or less invented retributivism as we know it in contemporary political philosophy.

In what follows we begin by describing the tenets in Anselm's theology that would later provide the materials for Kant's theory of retributive punishment. While Anselm, like Augustine before him, had included retributive elements in his theory of divine punishment, this did not imply that human beings were required to imitate those retributive elements in questions of temporal punishment. This "Augustinian asymmetry" allowed political mercy and divine retribution to coexist. We then discuss emerging trends in seventeenth and eighteenth century political thought that provided Kant with some of the materials needed to construct his influential retributivist theory. These thinkers undermined the foundations on which the Augustinian asymmetry

[1] This article draws on and reworks material from several chapters of our book *The Decline of Mercy in Public Life* (Cambridge: Cambridge University Press, 2014).

depended. Lastly, we show how Kant adapted Anselm's thought to reach political conclusions that Anselm's medieval and early modern successors had rejected. In making the above claims we draw on a variety of philosophical texts (examined more fully in our book) that highlight the historical path from Anselm's theology to Kant's retributivism.

Anselm's Theological Retributivism

Anselm's contribution to the philosophical position now known as retributivism—that is, punishing according to desert—drew on certain pre-existing features of western Christian theology, but also gave those features a distinctive and fateful twist. No pre-Christian philosophical or religious tradition had held as clearly retributivist a theory of divine justice as emerged in the western strain of Christian theology influenced by Augustine and his successors. They established clear agreement about the claim that God's final judgment would result in the eternal punishment of human sins. Just as significantly, however, throughout the early Christian period this retributive orientation did not lead Augustine and his successors to draw the conclusion that human justice should imitate divine justice by trying to apportion just deserts. Instead they drew the opposite conclusion: because God would one day take care of retributive justice far better than mere mortals could, human beings, faced with limited knowledge about human motives and desert, could focus instead on how to maintain order and bring their fellow citizens to repentance. These forward-looking goals were fraught with more than enough ambiguity and uncertainty to cause a ruler to lose sleep at night: the burdens of ensuring truly retributive responses to ill-desert, though expressly legitimized, were instead reserved to the omniscient heavenly Ruler who could observe and accurately judge not only external actions but also the heart within. Thus early Christian theories of divine justice did not tend to lead to retributive accounts of human punishment (understood as punishment according to desert), nor did they become a barrier to the display of mercy in temporal affairs. This position created a crucial asymmetry between divine and human justice which we call the Augustinian asymmetry.

Anselm consciously saw himself as a part of the larger Augustinian tradition. While there were clear retributive elements in Augustine's view of divine justice (infinite sin meriting infinite punishment, for example), Anselm's influential theory of the atonement clarified and amplified the tensions implicit in Augustine's theory between justice and mercy. The question of how far Anselm ought to be seen as the founder of a new retributivist impulse in western theology (and subsequently, western political theory) has become a subject of recurring scholarly discussion. For Gustaf Aulén, Anselm is the architect of the "Latin" view of atonement that diverges from the early Christian view.[2] Whereas early Christianity had emphasized the cross as the place where God defeated Satan, Anselm transformed it into the crucible where a wrathful God would be satisfied with nothing less than payment in full from humanity. On this interpretation, Anselm sees Christ as paying the debt as humanity's representative (or indeed as humanity incarnate), and so there is a sense in which humanity through Christ participates in its own salvation, paying the penalty in

[2] Gustaf Aulén, *Christus Victor: An Historical Study of the Three Main Types of the Idea of the Atonement* (London: SPCK, 1970).

full. This account focuses on God as the wrathful Father who demands repayment, rather than the fount of forgiveness and mercy. More recently, Timothy Gorringe sees in Anselm the roots of a retributive attitude toward temporal punishment that continues to the present day, a view that he also sees as out of step with earlier Christian thought.[3] Others argue that these interpreters misunderstand Anselm's position.[4] They suggest that Anselm did not in fact defend the doctrine of penal substitution (the claim that Jesus received the full punishment that sinful humans deserved) that has traditionally been attributed to him. Such interpretations, these critics say, ignore the fact that Anselm portrays God as undergoing extraordinary self-sacrifice to restore creation to its proper state. The former reading also represents Anselm's view of Jesus' death as mere retaliation, whereas it is better described as restitution. Our goal here is not to settle this interpretive question per se, but instead to assess the significance of Anselm's theology for the development of the concept of retribution in the west.

Anselm's distinctive contribution might therefore better be called, in an effort to bracket this question of penal substitution, the doctrine of mandatory satisfaction.[5] His argument centered on the claim that the wrong done to God's honor through human sin was so great that it would be unjust for God to simply forgive it. Instead, God had to seek payment from Christ on behalf of humanity to satisfy the demands of divine justice and honor.[6] Anselm's use of the concept of "satisfaction," so central to medieval theology, draws our attention to the debt metaphor because it is the language of restitution: someone has been wronged and is therefore owed compensation of some sort. Normally this metaphor emphasizes the freedom of the creditor (in the event of default) to choose whether to forgive the debt, to confiscate equivalent property, or to have the debtor punished in some way. In Anselm's theory, God's freedom to forgive is lessened because God, to act rightly, must uphold his own honor (because the order of the universe is harmed if God's honor is not upheld). As we will see, this claim establishes the first vital step in the creation of the claim that mercy violates the demands of retributive justice (we will refer to this as the

[3] Timothy Gorringe, *God's Just Vengeance: Crime, Violence, and the Rhetoric of the Salvation* (Cambridge: Cambridge University Press, 1996).

[4] D. Bentley Hart, "A Gift Exceeding Every Debt: An Eastern Otrhodox Appreciation of Anselm's *Cur Deus Homo*," *Pro Ecclesis* 7.3 (1993); George Huntston Williams, *Anselm: Communion and Atonement* (Saint Louis, MO: Concordia, 1960); Peter Schmiechen, *Saving Power: Theories of Atonement and Forms of the Church* (Grand Rapids, MI: W.B. Eerdmans Pub. Co., 2005); Harold J. Berman, *Law and Revolution: The Formation of the Western Legal Tradition* (Cambridge, MA: Harvard University Press, 1983), 179–191; John McIntyre, *St. Anselm and His Critics: A Re-Interpretation of the Cur Deus Homo* (Edinburgh: Oliver and Boyd, 1954).

[5] Among contemporary theologians, Lisa Sowle Cahill defends a form of substitutionary atonement. See Lisa Sowle Cahill, "Quaestio Disputata: The Atonement Paradigm: Does It Still Have Explanatory Value?" *Theological Studies* 68 (2007). For a defense of penal substitution more specifically, see Colin E. Gunton, *The Actuality of Atonement: A Study of Metaphor, Rationality and the Christian Tradition* (Grand Rapids, MI: W. B. Eerdmans, 1989); John E. Hare, *The Moral Gap: Kantian Ethics, Human Limits, and God's Assistance* (Oxford: Clarendon Press, 1996), ch. 10. Richard Swinburne accepts divine retributivism but rejects penal substitution. See Richard Swinburne, *Responsibility and Atonement* (Oxford: Clarendon Press, 1989). Timothy Gorringe critiques divine retributivism more broadly, interpreting the cross as God's negative judgment against the retributive orientation itself. See Gorringe.

[6] Some scholars note (correctly) that Anselm never actually affirms the doctrine of penal substitution, and that "satisfaction" can be thought of as a form of restitution rather than punishment. Nonetheless, it must be admitted that the "payments" Anselm had in mind did occur in the form of punishments—either the infinite punishment of the sinner to pay the infinite debt or Christ's death on the cross as a substitute payment.

mercy-justice paradox). Yet we will also see that for Anselm human mercy was not seen to be similarly paradoxical: his version of the problem was a theological rather than a political argument.

Anselm's theology led to a conception of mercy that was defined more directly in relation to justice, yet which eventually also created an influential perception that there is an inherent tension between mercy and justice. According to Anselm, God wills that some human beings be saved, yet God's justice will not allow God to save them by simply granting them mercy, because God's honor must be upheld (whether by punishment or by an alternative form of satisfaction). For Anselm, simple forgiveness (absent proper satisfaction) "is absolutely contrary to God's justice, which does not allow anything to be given in repayment for sin except punishment" (Anselm 1.24. p. 311).[7] Yet it is not ultimately Anselm's claims about sacrifice of Christ in our place that render the relationship between justice and mercy problematic; it is rather the claim that, if not for the provision of some form of satisfaction, divine mercy would have been unjust. Significantly, however, this pivotal theological conceptualization of divine justice took place without causing any significant disruption to Augustine's asymmetrical account of the relationship between justice in the divine and human spheres. Neither Anselm nor his immediate successors come close to claiming that temporal rulers who show mercy without demanding satisfaction are therefore unjust.

In *Why God Became Man*, Anselm begins from the Biblical teaching that God took on human nature and died to save human beings from their sin. He imagines people who sincerely puzzle about why God would have so acted "when he could have done this through the agency of some other person, angelic or human, or simply by willing it?" (Anselm 1.1, p. 265). Anselm's solution to this problem is to show that the incarnation and atonement of Jesus Christ were the *only* ways God's purposes could be fulfilled. To show this necessity, Anselm must refute the view that God could simply choose to forgive sin. If God could do so, then the Incarnation no longer seems necessary. Notably, Anselm's sketch of this problem closely mirrors key aspects of the traditional theological problem of evil. If God cannot forgive merely from his own will to forgive, we make God out to be powerless. If, however, God can forgive merely from his own will, but chooses not to, then God is either insufficiently wise or insufficiently benevolent. To defend God from these accusations, we need to show that God could not have saved man any other way than he did (Anselm 1.6, pp. 270–271).

What precludes the possibility of simple forgiveness is the importance in Anselm's theory of divine honor. In Anselm's understanding, we owe God all of our will, and failing to obey God accrues a debt, one increased by the additional insult inflicted on God's honor (Anselm 1.12, pp. 284–286). Under such circumstances, it is unfitting for God simply to forgive because doing so would leave sin wholly unregulated and unjustly treat the sinner and non-sinner the same. Anselm, like Augustine, envisions the drama of the atonement as centering on God's own justice and salvation, with the role of outside forces (such as the devil) decisively pushed to the margins. If

[7] All references to Anselm's texts are taken from Anselm of Canterbury, *The Major Works*, ed. Brian Davies and G. R. Evans (Oxford: Oxford University Press, 2008).

we ask why God teaches human beings to forgive when God himself does not simply forgive, the answer, according to Anselm, is that God has decreed that judgment and vengeance belong to him alone. In making this argument, Anselm approaches closest to applying his theological ideas directly to questions of political justice. Anselm begins by acknowledging the apparent contradiction: "[W]hen God teaches us to forgive those who sin against us," he notes, "he seems to be being contradictory—in teaching us to do something which it is not fitting for him to do himself" (Anselm 1.12, p. 285).[8] In response, Anselm appeals directly to Augustine's asymmetrical account with its distinction between the purposes of divine and human justice:

> There is no contradiction in this, because God is giving us this teaching in order that we should not presume to do something which belongs to God alone. For it belongs to no one to take vengeance, except to him who is Lord of all. I should explain that when earthly powers take action in this way in accordance with right, it is the Lord himself, by whom they have been appointed for the task, who is acting. (Anselm 1.12, p. 285)

Here we see that Anselm does not think human beings, in general, should imitate God by refusing to grant mercy without first obtaining satisfaction; instead, he believes the desire for avenging wrongs has no place among Christians, with one significant exception—the public authority. Like Augustine, Anselm assumes the legitimate authority of kings and princes.[9] Anselm argues that because temporal authorities are acting directly in the place of God, it is right for them to punish rather than forgive. Yet Anselm never draws the conclusion that mercy is wrong for public officials, and in other contexts he exhorts officials to be forgiving of wrongs against themselves.

In Anselm's view, it is an eternal, cosmic truth that God *necessarily* cannot lose his honor: human beings must honor God either voluntarily through an obedient will or involuntarily through punishment (Anselm 1.14. p. 287).

> If, however, God remits what he was about to take away from a person against his will, because of that person's incapacity to make payment—in that case he is making his punishment lax and making a person happy on account of his sin, in that the person has what he ought not to have. . . . But mercy of this kind is absolutely contrary to God's justice, which does not allow anything to be given in repayment for sin except punishment. (Anselm 1.24. p. 311)

[8] Here and elsewhere we cite language from the character "B" in the dialogue as speaking for Anselm. While it is clear that "A" is his primary spokesman, Anselm consistently employs in his dialogues an expository style of argument such that the statements of his interlocutors can always be treated as good faith concerns rather than intrinsically conflicting perspectives.

[9] Anselm interacted extensively with the king of England during his tenure as Archbishop of Canterbury. See Southern, *Saint Anselm*.

It is therefore "an absolute certainty, that God cannot remit a sin unpunished, without recompense, that is, without the voluntary paying off of a debt" (Anselm 1.19, p. 302).

Anselm now has the basic components in place to construct the paradox on which his argument rests. Humans owe God perfect obedience, and for obedience to be perfect it must proceed from the right motives. To obey God only because of other rewards and punishments is to dishonor him (Anselm, *On Truth* §12, p. 169). Any departure from this moral standard incurs an infinite debt, which human beings cannot repay because even perfect future obedience is no more than what is owed and, therefore, does not generate any surplus merit (Anselm 1.19–22, pp. 302–308). Since we already owe God perfect obedience, the best we can hope for is not adding to our debt: we can never repay it. Given the absolute dependence of the creature on the creator, Anselm states, "it is not sufficient merely to repay what has been taken away: rather, he ought to pay back more than he took, in proportion to the insult which he has inflicted" (Anselm 1.11, p. 283).

Thus far, however, we have only a very bad situation from humanity's perspective, not a full-blown paradox. The paradox emerges because God wills that human beings be saved (Anselm 1.16–18, pp. 289–300). Yet if human beings cannot pay the debt and God cannot simply forgive it, it seems God's plans will be thwarted, which Anselm believes is impossible. The heart of the paradox is that one of Adam's race must somehow make the payment for Adam's race, yet none of Adam's race can possibly accrue merit as a mere creature, and a sinful creature at that (Anselm 2.8, pp. 321–322). Christ, however, because he never sinned, and yet was tempted to sin and so accrued merit for his obedience, could through his divine nature offer something to God on behalf of humanity that was not already owed, namely a willingness to die innocently to restore God's honor.[10]

Anselm concludes that God is in fact merciful, so long as this mercy is understood in context of the mystery of the Trinity.

> What, indeed, can be conceived of more merciful than that God the Father should say to a sinner condemned to eternal torments and lacking any means of redeeming himself, "Take my only-begotten Son and give him on your behalf," and that the Son himself should say, "Take me and redeem yourself." (Anselm 2.20, p. 354)

Anselm thus reconciles God's justice and mercy, but is only able to do so by invoking two miracles: the existence of the Trinity and the incarnation of Christ.

For the purposes of our inquiry two facts seem evident. First, the moral order of the universe would, in Anselm's opinion, be compromised if God were to simply show mercy and refrain from sending sinful creatures to hell. Second, whether humanity is saved by an act of penal substitution or of restitution, the fact remains that something miraculous happened apart from

[10] Although Anselm's view is sometimes contrasted with the "moral influence" theory of Abelard, it is worth noting that Anselm also explicitly sees one purpose of the death of Christ to be moral example. See 2.18, pp. 348–352.

which God's mercy could not have been justly extended to human beings. This is important because Anselm's underlying concept of justice will eventually come to be appropriated by thinkers who no longer believe in the miracle of the Incarnation or in the reality of the atonement.

Thus Anselm's theology of the atonement does indeed represent a turning point in the development of retributivism the modern west, though not one obvious at the time or even in the centuries immediately following. Fundamentally, Anselm's version of the atonement introduced into western Christian thought the claim that it would be unfitting and unjust for God to simply forgive sins without receiving satisfaction. For Anselm and his contemporaries, however, the potential affinities between this theological argument about divine justice and a retributivist account of temporal punishment would not have been readily apparent. Indeed, from Anselm's point of view, the most striking thing about his account would not have been the vengefulness of a God who refuses to simply forgive, but rather the mercy of a God so determined to restore creation to its right state as to sacrifice his own Son. Nevertheless, the claim that sin necessarily requires satisfaction indicated a subtle shift in understanding that would carry lasting consequences for future western conceptions of justice and mercy. For this account of justice, shorn of its theological underpinnings, laid bare the inherent tension between mercy and justice implied by Anselm's theology, one that over the succeeding centuries would become a key element in the marginalization of public mercy.

Seventeenth and Eighteenth Century Roots of Kantian Retributivism

Although Anselm's successors, from Abelard and Aquinas to Luther and Calvin, departed from various aspects of Anselm's account of the atonement, as we argue in chapter four of *The Decline of Mercy*, they all retained his crucial premise that justice is the baseline against which mercy must be defined: mercy by definition is an act of liberality that goes beyond or even contradicts what justice requires. None of them would go so far as to say that justice is the opposite of mercy because they all believed that justice and mercy are essential attributes of God, making it problematic to call them opposites. From this point forward, however, western Christian thinkers typically conceptualized mercy as involving a choice to give more than is due, where what is due is determined directly by reference to an external standard of justice.

Seventeenth Century Natural Law's Rejection of the Augustinian Asymmetry: Grotius

In the early modern period the Augustinian asymmetry (positing less retributive standards for human justice than for divine justice) that had dominated the Christian era began to unravel, replaced by a growing preference for a more unified theory of justice and mercy. This resulted partly from theological changes—a growing skepticism regarding the mysteries of the Incarnation and atonement that had been central to Anselm's thought—and partly from political ones, notably the decline of monarchy and the corresponding rise of (more) egalitarian (quasi-) democracy, which helped replace deference toward monarchical judgment with an alternative moral framework that viewed discretion and even the potential for arbitrariness with profound suspicion and distrust.

One of the most notable changes to this consensus began to emerge within the modern natural law theories of the seventeenth century, in which we find the Augustinian asymmetry still lingering but beginning to erode significantly. In its place, we find an alternative approach to the relationship between divine and human punishment beginning to emerge—one that bases human punishment on a new understanding of the principles of natural law, and subtly imports those principles to justify God's punishments as well. These thinkers sought to create a more unified account by making divine justice less retributive, rather than by making human justice more retributive.

The major figures in this tradition, especially Hugo Grotius, but also to a lesser degree Thomas Hobbes and John Locke, developed their views about punishment and mercy in a context that raised significant questions about Anselm's account of divine retributivism and its relation to temporal punishment. One of their key interlocutors regarding these questions was the heterodox thinker Faustus Socinus, whose view of divine mercy captured in embryonic form some of the most novel elements of the views of political justice emerging in the seventeenth century. In Socinus's more individualistic world, the sins of a parent cannot properly carry over to the child, and each individual must be punished only for his own sin. Socinus was exploiting the tensions between Anselm's retributivist view of divine justice and the widespread Christian commitment to forgiveness that continued to legitimate at least occasional shows of mercy even among the cruellest of human rulers. Grotius and his successors answered this challenge in part by further marginalizing desert-based rationales for punishment as applied even to divine justice. For example, some (though not all) Socinians also challenged the traditional doctrine of hell and generally tended to reject the idea of everlasting torment.[11] Considering that one of the hallmarks of Socinianism was its insistence that what the Bible teaches (properly interpreted) must be reasonable, the fact that some Socinians rejected the traditional doctrine of hell is evidence that eternal retributive punishment no longer fit with the intuitive understandings of justice for an emerging segment of the early modern intellectual world.[12] Grotius (and later Hobbes as well) followed Socinus in rejecting a theology that interprets God as a judge bound by retributivist logic who lacks the ability to pardon; his fundamental premise is that God must instead be understood as a sovereign ruler positioned above the law rather than constrained by it (Grotius DSC 2, p. 53).[13]

[11] John Biddle, the "Father of English Socinianism," argued in his *Twofold Catechism* that hell was merely an everlasting death, not a place of eternal torture. John Biddle, *A Twofold Catechism : The One Simply Called a Scripture-Catechism: The Other a Brief Scripture-Catechism for Children* (London: Printed by J. Cottrel for R. Moone, 1654), 135–138. See also the discussion of Hobbes later in the chapter.

[12] For more discussion, see H. John McLachlan, *Socinianism in Seventeenth-Century England* (London: Oxford University Press, 1951). For a general discussion of views on hell in seventeenth-century England, see Philip C. Almond, *Heaven and Hell in Enlightenment England* (Cambridge: Cambridge University Press, 1994).

[13] Grotius references "DSC" refer to Hugo Grotius, *A Defence of the Catholic Faith Concerning the Satisfaction of Christ, against Faustus Socinus* (Andover: W. F. Draper, 1889. References to RWP refer to Hugo Grotius, *The Rights of War and Peace* (Indianapolis: Liberty Fund, 2005).

Alongside this challenge was another change of immense importance: the increasing concern about mercy's arbitrariness, stemming from the growing emphasis in the seventeenth century on political equality. While not all modern natural law thinkers endorsed a normatively robust conception of equality, they affirmed at a minimum the importance of equal treatment before the law as a strategy for maintaining peace. Given that historically mercy had been seen as highlighting the inequality between the one bestowing and the one receiving mercy, this full-on assault against arbitrary rule had collateral consequences that contributed directly to mercy's marginalization.

Based on his theory of natural law, Grotius views punishment as morally justified even prior to the foundation of society. On this theory, punishment should generally be viewed as a right rather than a duty: punishment may sometimes be a moral necessity (as in the case of the atonement), but only because a pressing moral end (such as maintaining good order) calls for it, and not intrinsically. Grotius's theory of punishment takes to the next level the generally forward-looking orientation of the Christian-era approach by systematically excluding non-forward-looking considerations. Indeed, one of the most notable developments in the modern natural law tradition is how readily these thinkers adopted the forward-looking rationale as the preferred approach to justice in *both* the human and divine spheres. Grotius's account of the atonement contributes to this substantially by portraying God as a forward-looking governor, and Grotius derives from that account a corresponding rationale justifying human punishment. Yet Grotius stops short of his seventeenth-century successors in one important respect: he continues to claim that God is not required to have forward-looking justifications for punishment and that God at times punishes simply from a desire to punish evil (Grotius RWP 2.20.4, pp. 956–958). Grotius thus still acknowledges the validity of divine punishment that remains simply retributive in nature.[14] Yet he is adamant that the situation of human beings is not analogous, and that consequently human beings must only use forward-looking rationales.

This doubling down on the forward-looking human justifications embraced by earlier Christian thinkers had the effect of focusing questions about mercy exclusively on the issue of future consequences. It is against this background that Grotius understands the underlying rationale for justified clemency (Grotius RWP 2.20.22, p. 997). Here Grotius offers perhaps the ultimate reason for his hesitations about retributivism: for any law, there always remains a lawgiver who retains authority over that law and can adjust it to better suit the lawgiver's purposes. This claim had already featured in Grotius's portrayal of God as ruler rather than injured party in his discussion of the atonement, denying specifically "that criminals are punished by the prince because they injure the state, of which he is the head." On the contrary, he asserts,

[14] On the basis of this passage, De Pauley wants to characterize Grotius's punishment as primarily retributive, with utilitarian constraints. This seems to us backwards. For human punishment, retributive justifications are practically unimportant, and for divine punishment, God is portrayed as normally operating on the basis of forward-looking reasons even though God is not unjust in punishing retributively. See W. C. De Pauley, *Punishment Human and Divine* (London: SPCK, 1925), 116–118.

> this right belongs to the ruler as ruler. As soon as you establish supreme power, you establish the right of punishing. Take away the one, and you take away the other. Whatever is said of the right of punishing must necessarily be understood of the right of forgiving. (Grotius DSC 2, p. 57)

Despite this authority, however, Grotius maintains that the lawmaker "ought not to take away the Law, without a reasonable Cause for it, which if he does, he transgresses the Rules of political Justice" (Grotius RWP 2.20.24.1, p. 999). As we have seen, this precisely mirrors the rationale for God's mercy: his ability to pardon is not unrestrained, but the restraints that do apply come from the logical implications of the kind of creation God as ruler wills to sustain (Grotius RWP 2.20.36, pp. 1016–1017). Here we see the beginnings of the view that Kant will eventually develop, in which mercy is ruled out, not for consequentialist reasons, but rather for the sake of freedom and equality.

The Return of Retributivism

By the end of the seventeenth century, the modern natural law tradition had decisively influenced the direction taken by western conceptualizations of both punishment and mercy. Throughout the Christian era up to the Protestant Reformation, the instinctive consensus of western thinkers had centered on Augustine's asymmetric account of divine and human justice: a blending of a substantially retributive conception of divine punishment and a grace-oriented conception of divine mercy juxtaposed in the temporal sphere against primarily forward-looking rationales for both punishment and mercy. The modern natural law theorists, as we have seen, modified this consensus substantially. They offered instead a thoroughgoing rejection of retributive rationales as well as a harder-line insistence on the exclusivity of forward-looking justifications for punishment—and, implicitly, for mercy as well.

Over the course of the development of eighteenth century moral and political philosophy, however, a revived form of retributivism emerged among thinkers such as Samuel Clarke, Richard Price, Jean-Jacques Rousseau, and Adam Smith. Their counter-move involved accepting in part the central premise of the modern natural law era critics of retributivism—namely, the doubts they had raised about the asymmetric requirements of divine and human justice and their compatibility with the increasingly valued ideas of universalizability and impartiality as moral criteria. Instead, these thinkers separately introduced elements of a non-asymmetric retributivism that would eventually merge in the form of a kind of secularized Anselmian theology in the thought of Immanuel Kant. These pre-Kantian thinkers shared a focus on the importance of providing an intrinsically fitting response to wrongdoing, as well as a deepened commitment to equality and impartiality as important moral values in their own right.

Moral Realism—Clarke and Price

The triumph of the forward-looking rationale for punishment over retributivism in the seventeenth-century modern natural law tradition did not lead immediately to a similar retreat of retributive

rationales for divine punishment. On the contrary, theories positing that God punishes sinners eternally and retributively continued to dominate theological writing throughout the Enlightenment, and these theories inevitably had a significant influence on the rationales offered for earthly punishments as well. We find this impulse at the heart of another important tradition of moral philosophy during the period, one that we call "moral realism" (borrowing a term from contemporary philosophy).[15] This designation highlights the fact that these thinkers believed in an inherent rightness and wrongness, fitness and unfitness, associated with actions that could be perceived in the very structure of the universe rather than constructed from the reactive attitudes of human beings.[16] Moral realism in this sense had a natural affinity with retributivism, as retributivists tended to rely on the perception that there is always a proper and fitting punishment applicable to a given situation. If our sense of retributive justice tracks with an objective morality, this substantially deflects the charge that retribution is merely revenge writ large. One early proponent of moral realism, Samuel Clarke, and one later proponent, Richard Price, together exemplify the moral realists' attempt to harmonize traditional Christian ethical precepts with arguments that could be supported by reason alone.

Clarke claims that certain necessary and eternal relations between phenomena carry with them a set of universally valid implications about the "consequent Fitness or Unfitness" of actions (Clarke 45).[17] Such moral principles derive from "the eternal differences of Good and Evil" and "the unalterable rule of Right and Equity" and display a logic irresistible to both man and God. Thus, according to Clarke, eternal moral truths known to natural reason—it is clear in context that he has in mind natural law—must by virtue of their eternality be antecedent to, and thus also logically prior to, any rewards and punishments that may later attach to them. What instead creates the real moral obligation is "right reason," understood as being equivalent to "impartial reason" (Clarke 88, 113). Impartial reason provides its own sanction in the form of an authoritative conscience that makes those who propose to violate right "a Law unto themselves" (a formulation that was later to become central to the deontological tradition of ethics) (Clarke 189; see further 71). Our impartiality in cases in which we are disinterested proves we naturally apprehend "the unalterable difference of Right and Wrong" (Clarke 75).

Moral realism undergirds Clarke's account of divine retributive punishment. When God punishes, Clarke observes, we may rely on it that "the Punishment which shall be inflicted on the Impenitent, shall be exactly proportionate to their Sins, as a recompense of their demerits, so that no man shall suffer more than he has deserved" (Clarke 335). Nevertheless, Clarke takes this account of divine retribution to be compatible with at least some version of divine mercy. Most of

[15] Christine Korsgaard, in *The Sources of Normativity*, precedes us in using the term "moral realism" to designate this tradition of thinkers. Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996). See also Paul Formosa, "Was Kant a Moral Constructivist or a Moral Realist?," *European Journal of Philosophy* (2011).

[16] It is tempting to call them "moral intuitionists," yet they do not consistently appeal (as Hutcheson and others had) to a special "moral sense" that directly perceives these moral relations.

[17] Parenthetical references to Clarke are taken from Samuel Clarke, *A Demonstration of the Being and Attributes of God; A Discourse concerning the Unchangeable Obligations of Natural Religion* (Stuttgart-Bad Cannstatt: Friedrich Frommann Verlag, 1964).

our knowledge of God's mercy we derive from revelation rather than natural reason. From the natural law we can infer that divine pardon for sin is a possibility but not that it is a certainty. We are equally led to infer that God will likely require satisfaction or expiation in order to vindicate his honor and provide "sufficient Testimony of his irreconcilable Hatred" against sin (Clarke 224–225, 309–312). Clarke also holds that this method of atonement – expiation plus forgiveness – has the effect not just of satisfying, but maximizing God's honor. Through Christ's atonement God is able to receive both the honor due his magnanimity in offering universal forgiveness as well as the honor due his authority in inflicting punishment for sin on the sacrificial victim (Clarke 224–225, 309–312). Thus, in Clarke's framework, there remains room for a form of divine mercy even beside a robust moral realism and a strong direct relationship between desert and punishment.

In Price we see a later and more developed version of Clarke's approach by a contemporary of Adam Smith. Price's theory follows Clarke's in seeing moral truths as being inherent in the very nature and relations of metaphysical reality. More directly than Clarke, however, Price frames his moral realism specifically in terms of good and ill-desert, with the meaning of desert articulated directly in terms of fitness for punishment. Just as the fitness of an action is the same as its rightness or wrongness, so the fitness of an agent is her good or ill-desert, which, like rightness, Price takes to be a primitive and natural category (Price 4, p. 79).[18] To say an agent is deserving is to say that rewarding him is proper and punishing him improper; likewise, to call an agent undeserving signifies that punishment is proper and reward improper (Price 4, p. 79).

This bears directly on Price's account of punishment. In contrast to the purely forward-looking account given by the natural law tradition, Price contends that future considerations alone cannot possibly justify punishment. The reason we approve of punishing someone is not that doing so serves that person's own happiness or that of others, since if this were so, we would be morally indifferent between two equally happy persons, one of whom has achieved good character and the other of whom has continued to be vicious (Price 4, p. 82). Obviously we do not approve these two alternatives equally, despite the fact that the net happiness produced by them is (by hypothesis) the same. The reason we do not, Price maintains, is that punishment is to some degree inherently fitting as the appropriate response to vice and not just in consequence of its tendency to deter, protect, or rehabilitate. It is this direct apprehension of good and ill-desert that grounds "the passion of resentment," to which Price gives a special role in ensuring that questions of justice receive due consideration (Price 4, p. 83).

Clarke and Price thus began to develop an account of punishment as an inherently fitting response to ill-desert independent of any consequences it might produce in social life. With Clarke and Price, this account could be reconciled with a fairly liberal exercise of both divine and human mercy. As we will see, however, it later came to provide a foundation for the revival of the more thoroughgoing retributivism we encounter in Smith and in Kant.

[18] Parenthetical references to Price refer to Richard Price, *A Review of the Principal Questions in Morals*. trans. D. D. Raphael (Oxford: Clarendon Press, 1974). (By chapter and page number.)

Impartiality—Adam Smith

There are two main lines of thought that may lead us to regard mercy as unjust: that it fails to give what is due (moral realism) or that it fails to treat like cases alike (impartiality). Thus far we have seen in thinkers like Clarke and Price a commitment to retributive punishment primarily on grounds that it is inherently fitting. In the thought of Smith, the emphasis shifts to the importance of impartiality.

We have seen how moral realists such as Clarke and Price affirmed retributive punishment in a less overtly theistic way by interpreting certain actions as intrinsically deserving of reward or punishment. David Hume had criticized this approach, arguing that at base our moral judgments reflect our desires and passions. Adam Smith sought to reconcile these two positions. Like Hume, he affirmed sentiment as the basis for moral judgments, yet he rejected the idea that our moral judgments do no more than track social usefulness. Instead, Smith developed an alternative approach in which our sentiments are moral if they correspond with the sentiments of an impartial spectator. This allowed him to affirm retribution (as Clarke and Price had) without reducing our reasons for so affirming it to ones of mere utility. One way of reading Kant, to whom we will turn in the next section, is that he recasts Smith's impartial spectator in terms of reason rather than sentiment (and legislating instead of spectating), and within this new framework provides a rationale for retributive punishment that does not rest on forward-looking justifications. By contrast, Smith's sentimental version, as we will see, remained at least somewhat more hospitable to mercy than Kant's rationalist alternative.

The idea of the impartial spectator is not original to Smith: both Joseph Butler and Price had developed versions of it. However, Smith's impartial spectator is the most complex and substantive conception of reflective endorsement prior to Kant. Moreover, Smith makes clear that in the context of earthly ethical questions, the impartial spectator is to be regarded as serving the same function that God had served in more voluntarist divine command theories.[19] The spectator's imagined judgments, according to Smith, are "to be regarded as" the commands of God, "promulgated by those vicegerents which he has thus set up within us," and enforced by the punishment of self-disapproval (Smith TMS III.5.6, pp. 165–166).[20] Given that punishment is the appropriate response to resentment, it follows that moral ill-desert is the behavior that appears to the impartial spectator to be the proper object of punishment—the object of our "indirect sympathy with the resentment of the sufferer" (Smith TMS II.i.1.3, p. 68; II.i.2.1–2, p. 69; II.i.5.4–5, p. 75).[21]

[19] Parenthetical references to Smith TMS are taken from Adam Smith, *The Theory of Moral Sentiments*, ed. D. D. Raphael and P. G. Stein (Indianapolis: Liberty Classics, 1982). (By book, chapter, section, part and then page number.) References to Smith J refer to Adam Smith, *Lectures on Jurisprudence*. ed. R. L. Meek, D. D. Raphael, and P. G. Stein (Indianapolis: Liberty Classics, 1982). (By page number.) Smith says the origin of primitive religions was the invention of gods who were taken to represent and to enforce the judgments of an impartial spectator (Smith TMS III.5.4, p. 164).

[20] At Smith TMS III.2.31, p. 128, Smith also refers to human judgment as God's "vicegerent" and "substitute" on earth and as "the tribunal established in their own breasts" that is "superior" to any other earthly authority, though this passage was removed in later editions.

[21] See also Smith TMS II.i.4.3–4, p. 74; II.ii.1.2, p. 78, II.ii.2.1, p. 82; VI.ii.intro.2, p. 218. See also Vivienne Brown, *Adam Smith's Discourse: Canonicity, Commerce, and Conscience* (London: Routledge, 1994), 120.

Smith's belief that punishment properly expresses the impartial spectator's sympathy with the resentment of the injured party causes him to deny emphatically that punishment is grounded in public utility. It also leads him toward a rehabilitation of retributivism through a reflective endorsement-based account of the necessity of punishment (Smith J, p. 475). Smith favors retributive language more than most of his Enlightenment predecessors. He calls retaliation "the great law which is dictated to us by Nature" and defines punishment as "to recompense, to remunerate . . . to return evil for evil that has been done" (Smith TMS II.i.1.2, p. 68). Smith's theory of punishment thus constitutes a nearly comprehensive restoration of the retributive outlook, and indeed he is the first major modern philosopher to explicitly challenge the forward-looking rationale of Grotius and his followers (Smith J, p. 136). As we have seen, Smith maintains that pursuing punishment understood as "the revenge of the injured" will in practice get us all the forward-looking benefits (protection, deterrence, restitution, etc.), which Grotius and his followers had praised (Smith J, p. 105). However, that is neither the true origin of punishment nor an ethically permissible rationale for it (Smith J, p. 104).

Along these same lines, Smith precedes Kant in characterizing the death penalty in cases of murder as non-optional, a claim he premises directly on the viewpoint of the impartial spectator (Smith J, pp. 104–106). Smith's crucial assumption is that in punishment-related contexts, the magistrate stands in for the impartial spectator. Here Smith follows most of the modern natural law theorists in stressing the sovereign's neutral position between his subjects as providing him proper standing to punish offenders in civil society. Smith's retributivism, however, does not lead him into an unqualified hostility to mercy. Partly, this residual openness to mercy comes from grounding his retributivism in the virtuousness of our moral sentiments, particularly sympathy, which obliges him to praise the sentiments inclining toward mercy. Contrary to Kant as well, Smith is willing to allow forward-looking considerations to factor into the question of what the appropriate proportion is between punishment and crime. It is of note, however, that Smith is able to move this far in the retributivist direction while still grounding his system directly on sympathy for others. It is easy to see how, once sympathy is denied a legitimate place in strictly moral calculation, as in Kant's ethics, the path to a more unyielding version of retributivism is short and uncomplicated.

Finally, Smith's views on divine justice also have implications for retributive justice. When Smith speaks about divine justice, it corresponds to his broader retributive view in most respects: he characterizes the general rules of morality as laws of God, the violation of which is and should be punishable in the afterlife. Indeed, Smith places considerable importance on the belief that God will ultimately judge all actions rightly as a reason for avoiding despair at the injustices of the here and now (Smith TMS III.2.33, pp. 131–132). Yet Smith never quite says that there is such a God, only that it is beneficial to believe it (Smith TMS VII.iv.11, p. 331). This evasiveness is not an isolated case. When Smith addresses in passing the doctrine of the atonement, he similarly assumes a generally retributivist account of divine punishment, including an Anselmian interpretation of

the significance of the crucifixion (Smith TMS II.iii.intro.2, p. 92).[22] However, in the final edition of TMS published during his lifetime, Smith cut the paragraph addressing the atonement, and there has been controversy ever since about how much to read into that excision.[23] One reading of these ambiguities is that Smith recognized the social value of belief in a God who would ultimately reward and punish with perfect retributive justice, but over time he came increasingly to doubt whether this belief was true. From Augustine to Calvin, belief in a God who would reward and punish retributively on the last day was thought to relieve human rulers of much of their duty to ensure that offenders received their just deserts through temporal punishments. However, as confidence in that day of future judgment receded, those like Smith who were committed to retributive justice as a moral principle began to see the full weight of that demand as necessarily falling instead on human shoulders, bringing human mercy into decisive tension with the demands of human retributive justice.

What Clarke, Price, and Smith share in common is a desire to detach, at least to some degree, our evaluative judgments about agents themselves from our judgments regarding the consequences of their actions, a desire to focus instead on their acts and the states of mind giving rise to them by inquiring about their intrinsic fittingness. Many of these arguments were framed primarily in the familiar terms of eighteenth-century moral philosophy—rightness and wrongness, propriety and impropriety, merit and demerit—and they invariably used fitness for reward or punishment as the key for recognizing rightness and wrongness. Early versions of this view, such as those found in the writings of Clarke and Price, grounded such judgments in a foundational moral realism, which in turn oriented these thinkers toward a more essentialist account of punishment substantially at odds with the forward-looking view. Yet these moral realist theories proved vulnerable, particularly amid the secularizing currents of the times, precisely based on their foundationalist character: the irreducible role they were supposed to play within the moral system created an inability to further justify their claims.

Adam Smith developed an alternative account resistant to these vulnerabilities by grounding moral judgments in our reflective endorsement of actions. This allowed retributivism to retain its absolutist character without having direct recourse to the theological underpinnings so conspicuous in the theories of Clarke and Price. Smith's theory also paved the way for Kant's further development of a reflective endorsement account of morality, which, as we will see, gave rise to a particularly unyielding retributivism when applied to the problem of punishment. Over the course of the eighteenth century, this preoccupation merged with a related sea change within political discourse: the emergence of generality, universality, and equality as central commitments of the Enlightenment political theories. The effect of this combination was to bring impartiality, already a key concept in the seventeenth-century modern natural law theories, to the very center of modern debates about punishment and mercy. These trends converged dramatically in Smith's impartial spectator, a device that could ground a strongly retributive approach to punishment and

[22] See further Smith, TMS, Appendix II, pp. 383–401.
[23] Similarly, while Smith's first lectures were on natural theology, they were destroyed before his death, likely at his request. See further K. Kim, "Adam Smith: Natural Theology and Its Implications for His Method of Social Inquiry," *Review of Social Economy* 55.3 (1997).

that also helped lay the groundwork for to the more lastingly influential version of impartiality we encounter in the philosophy of Kant.

Kant's Secularization of Anselm

Kant famously staked out perhaps the most extreme retributivist position in western philosophical ethics. Kant wrote in *The Metaphysics of Morals*:

> The law of punishment is a categorical imperative, and woe to him who crawls through the windings of eudaimonism in order to discover something that releases the criminal from punishment or even reduces its amount by the advantages it promises. (Kant MM 6:331)[24]

For Kant, public mercy in the form of a pardon or reprieve is inherently unjust; we must instead do whatever justice requires irrespective of the consequences for human happiness. Infamously, Kant goes on to say that

> even if a civil society were to be dissolved by the consent of all its members . . . the last murderer remaining in prison would first have to be executed, so that each has done to him what his deeds deserve and blood guilt does not cling to the people for not having insisted upon this punishment; for otherwise the people can be regarded as collaborators in this public violation of justice. (Kant MM 6:333).

Mercy obviously cannot be a virtue within this Kantian framework.

We argue in chapter eight of our book that the situation is more complex than these famous passages would indicate, as shown by the fact that Kant was willing to affirm mercy by private citizens and even by kings and princes in cases where the crimes committed were against the ruler, as in the case of civil war (Kant MM 6:334-7 and 6:460-1).[25] Instead, we argue that Kant's strongly retributivist account is motivated primarily by two general features of his ethical theory. First, Kant's commitment to equal respect for persons made him particularly concerned with impartiality in the administration of public justice, treating like cases alike. Thus in a famous passage from *Perpetual Peace*, Kant argued that

[24] Parenthetical references to Kant LE refer to Immanuel Kant, *Lectures on Ethics*, trans. Peter Heath, ed. Peter Heath and J. B. Scnheewind (Cambridge: Cambridge University Press, 1997). (By page numbers corresponding to those in the Royal Prussian Academy edition.) References to Kant MM refer to Immanuel Kant, *The Metaphysics of Morals*, trans. Mary J. Gregor (Cambridge: Cambridge University Press, 1996). (By page numbers corresponding to those in the Royal Prussian Academy edition.) References to Kant R refer to Immanuel Kant, *Religion and Rational Theology*, trans. and ed. Allen W. Wood and George di Giovanni (Cambridge: Cambridge University Press, 1996). (By page numbers corresponding to those in the Royal Prussian Academy edition.)

[25] See also Kant LE 27:688-90.

> The proverbial saying *fiat iustitia pereat mundus*[26] (i.e., let justice reign, even if all the rogues in the world must perish) may sound somewhat inflated, but it is nonetheless true. It is a sound principle of right, which blocks up all the devious paths followed by cunning or violence. But it must not be misunderstood, or taken, for example, as a permit to apply one's own rights with the utmost rigour (which would conflict with ethical duty), but should be seen as an obligation of those in power not to deny or detract from the rights of anyone out of disfavour or sympathy for others.[27]

If Kant's primary concern were ensuring that offenders received their retributive deserts without exception, we would expect him to say each wronged person has a duty to press for the full extent of punishment. Instead, it is only rulers and magistrates who are obligated to impose the utmost rigor of the law, while mercy on the part of private individuals he finds not only permissible but frequently laudable.

Second, Kant's uncompromising rejection of consequentialist moral reasoning led him to reject any view of punishment with even a hint of instrumentalism (Kant MM 6:215, 6:331). If considerations about consequences are permitted to infect our judgments about happiness, we can no longer have confidence in their strict conformity with principles of right—principles that, for Kant, must be universalizable in order to retain their distinctively rational character. This motivates Kant's conclusion that for any crime "only the *law of retribution* (*ius talionis*) . . . can specify definitely the quality and the quantity of punishment" (Kant MM 6:332). Retaliation "is by its form always the principle for the right to punish" because it alone can be an *a priori* rather than conditioned and conditional criterion (Kant MM 6:363). If we punish a criminal for purposes of deterrence, he may complain that we are using him as a mere means to an end and therefore not respecting his intrinsic personhood; he cannot make the same complaint about the retaliatory remedy, because in that case "he brings his misdeed back upon himself" and experiences only "what he has perpetrated on others" (Kant MM 6:363).

However, as we acknowledge in our book, there is more to Kant's retributivism than his commitment to impartiality and his rejection of eudaimonism. In addition to these key concerns drawn from his larger philosophical commitments, Kant clearly thinks there is something intrinsically fitting about the punishment matching the crime. This indicates that for Kant punishment according to desert is valuable for its own sake, and not just because it is necessary to uphold impartiality and rational consistency. We can understand this aspect of Kant's retributivism as reflecting Kant's decision to retain certain Anselmian assumptions about the nature of justice,

[26] Kant here aligns himself with the strictly retributivist policy of Frederick William I and implicitly opposes the reforms of Frederick William II (Frederick the Great), who opposed many of his father's policies on utilitarian grounds and worked to apply enlightenment principles to the criminal law. Robert Reinhold Ergang, *The Potsdam Führer: Frederick William I, Father of Prussian Militarism* (New York: Octagon Books, 1972), 134; Richard J. Evans, *Rituals of Retribution: Capital Punishment in Germany, 1600–1987* (Oxford: Oxford University Press, 1996), 121–140.
[27] Immanuel Kant, *Kant: Political Writings*, ed. H. S. Reiss, trans. H. B. Nisbet, 2nd, enl. ed. (Cambridge: Cambridge University Press, 1991), 123.

while simultaneously rejecting much of the traditional Christian theological framework that had supported it.

Anselm, as described above, had developed an account of divine justice in which human beings have an obligation to obey God perfectly, yet where there is no way they can repay God for even a single sin (because any future obedience only gives God what he was already owed). God's justice will not allow God's honor to be damaged by simply forgiving the debt, even though he permits Jesus to pay the debt on humanity's behalf. If one were to strictly apply this logic to human justice (without the intervention of the atonement), the implication would be that mercy inherently contradicts the requirements of retributive justice. However, as noted above, Christian thinkers generally did not feel obligated to draw this conclusion, understanding divine and human punishment to rely on two separate rationales.

Kant's retributivism can be described as a vestige of Anselm's account of justice, but with two crucial changes. First, in Kant's thought, the asymmetrical rationales for divine and human justice disappear, requiring political authorities to use the same uncompromising retributivist logic Anselm had attributed to God. Kant is insistent that the principles of justice are the same for all rational beings.[28] From this perspective it is impossible to argue, as Christians in previous centuries had, that human punishment might rely partly on consequentialist reasoning while divine justice remained retributive in character. Second, Kant's thoroughly individualist ethical theory jettisoned crucial elements of the traditional atonement narrative, making it incoherent to claim that Jesus, or anyone else, might pay another's moral debt. With these changes in Kant's underlying assumptions, public mercy began to seem an unjust way of using discretionary power, regardless of the fairness of the procedures by which it might be applied or the consequences to which it might lead.

We begin by noting this pattern in Kant's conception of God. The notion that Jesus suffered punishment in place of sinners provides a means by which both God's justice and his mercy may be satisfied at once. Kant's retributivism, however, stresses justice over mercy to such an extent that Kant's God seems unable to accommodate even this much of a departure from the strict application of the *lex talionis*. Kant explicitly denies that we should represent God "as *merciful* and hence *forbearing* (indulgent) toward human weakness" and maintains that God's justice "cannot be represented as *generous* and *condoning* (for this implies a contradiction)" (Kant R 6:141). This last point Kant repeats elsewhere by observing that the idea of a "generous judge" is in fact a contradiction in terms, because if we presuppose the judge's generosity, we have already impeached the judge's impartiality (Kant R 6:141).

In this respect, Kant's distinctive move is to retain Anselm's theological retributivism while abandoning the characterization of Christ's atonement as able to effect the necessary

[28] Immanuel Kant, *Groundwork of the Metatphysics of Morals*, trans. and ed. H. J. Paton (New York: Harper and Row, 1964), 4:389.

satisfaction.[29] Here we observe the close linkage between Kant's belief in impartiality and his more foundational commitment to the principle of universalizability. "So far as we can judge by our reason's standards of right," he states, "this original debt [of sin] . . . cannot be erased by somebody else," because "it is not a *transmissible* liability which can be made over to somebody else, in the manner of a financial debt (where it is all the same to the creditor whether the debtor himself pays up, or somebody else for him)" (Kant R 6:72). Instead, punishment involves "the *most personal* of all liabilities, namely a debt of sins which only the culprit, not the innocent, can bear, however magnanimous the innocent might be in wanting to take the debt upon himself for the other" (Kant R 6:72). In the end, Kant insists that the idea of God *pardoning* sin in the human sense of the term is simply unthinkable. God must instead "allow each only that measure of happiness which is proportionate to his worthiness," or else implicitly forfeit God's own moral integrity (Kant R 28:1086–87). Kant thus sees the need for satisfaction but does not believe a third party can make satisfaction on behalf of someone else: given that sin is fundamentally a function of an individual's wrong will, liability for sin is for that very reason essentially nontransferable. Kant also denies that a person who ceases to do wrong can in any sense be understood to erase the wrongs previously committed. This is significant because for Kant, any violation of the moral law carries with it "an *infinity* of guilt," and therefore also "*infinite* punishment and exclusion from the Kingdom of God" (Kant R 6:72). Kant also makes clear that for this same reason, only punishments—never rewards—are, strictly speaking, the subject of justice. "Even if we unceasingly observe all moral laws, we can never do more than is our duty," Kant argues, following Anselm's argument step by step. Consequently, "we can never expect rewards from God's justice," only punishment (Kant R 28:1085).

In his *Lectures on the Philosophical Doctrine of Religion*, Kant offers a straightforward formula for understanding God's justice, arguing that "*limitation of benevolence by holiness* in apportioning happiness is *justice*" (Kant R 28:1074). By analogy, justice places similar limits on benevolence in the case of human justice as well. For the human judge faces the same dilemma as God himself:

> I must not think of a judge as benevolent, as if he could somewhat relax the holiness of the law and spare something of it. For then he would not be a judge at all, since a judge must weigh and apportion happiness strictly according to the measure in which the subject has become worthy of it through his good conduct. The justice of the judgment must be unexceptionable and unrelenting. (Kant R 28:1074)

Once one has defined justice with the precision Kant does—and on the analogy with divine justice, of course, no degree of precision would be excessive—there remains no unoccupied moral space in which benevolence, or mercy understood in terms of benevolence, might be expected to play a significant role.

[29] In *The Metaphysics of Morals*, Kant does make a remark that sounds like he is drawing on the substitutionary atonement theory in an approving sense (Kant MM 6:490). Elsewhere, however, Kant argues against the possibility of a genuine substitutionary atonement in no uncertain terms (Kant R 6:72, 28:1087).

Kant thus arrives at a similar paradox to Anselm's, only with the crucial theological foundations of the latter removed. Human beings must be perfectly obedient to the moral law; they must will what is right for its own sake and not for the sake of rewards and punishments. Given that the moral standard is perfection, there is no way for present or future obedience to adequately compensate for past disobedience, because even perfect obedience is no more than merely what was owed. Yet Kant's individualist ethics excludes as an alternative the traditional Christian doctrine of Christ's vicarious sacrifice, leading Kant to formulate instead one of the oddest, and least merciful, theories of the atonement in Christian history.[30] Although Kant's conception of justice closely mirrors the retributivist aspects of Anselm's theology in many respects, he also adapts a key feature of Peter Abelard's critique, namely Abelard's rejection of the idea that Christ' death could be efficacious in making satisfaction for an *unrepentant* sinner, and his insistence instead that God's forgiveness is inherently conditional on a change of heart in the offender (Kant 4 28:1084). Like Abelard, Kant argues in *Religion within the Bounds of Mere Reason* that a person who wishes to become worthy of happiness "must make or have made *himself* into whatever he is or should become in a moral sense, good or evil" by renouncing and repenting of the original sin of Kantian morality: letting one's conduct be determined by any sort of temporal incentives (Kant R 6:44). The human being, Kant explains, becomes good or evil "according as he either incorporates or does not incorporate into his maxims the incentives contained in that predisposition (and this must be left entirely to his free choice)" (Kant R 6:44). In this way, the act of moral conversion itself becomes a kind of "*punishment* whereby satisfaction is rendered to divine justice" (Kant R 6:74). Moral conversion in this comprehensive sense involves "the death of the old man," and "the crucifying of the flesh," by means of which "the new human being undertakes [suffering] in the disposition of the Son of God, that is, simply for the sake of the good" (Kant R 6:74). This reference to "the disposition of the Son of God," moreover, telegraphs Kant's radically unorthodox strategy for addressing the vicarious substitute end of the atonement equation. Instead of Jesus making satisfaction for the sinner, in Kant's theory the sinner makes satisfaction *for himself* by becoming a new being with a disposition like that of Jesus, punishing the self as a means of converting to the sought-after new disposition. More precisely, in Kant's story it is our future perfected self that makes satisfaction for our present sinful self (Kant R 6:47, 72–76, 176–178, 137–146).[31]

However, the nature of the change of heart that constitutes these sufferings, it soon becomes clear, is a vision of moral conversion such as only Kant could have conceived: it is a

[30] John Hare, *The Moral Gap: Kantian Ethics, Human Limits, and God's Assistance* (Oxford: Clarendon, 1996), 53–68, offers an astute analysis.

[31] Kant calls "*grace"* God's "imputation" to us of the righteousness of the vicarious substitute of the perfect future man. God alone perceives our future trajectory and therefore imparts to us this righteousness even when we have not yet earned it. He also asserts that it is, however, "fully in accord with eternal justice" for him to do so because of the "satisfaction" created by our change of heart (Kant R 6:76). See also Colin E. Gunton, *The Actuality of Atonement: A Study of Metaphor, Rationality and the Christian Tradition* (Grand Rapids, MI: W. B. Eerdmans, 1989), 7; David Sussman, "Kantian Forgiveness," Kant-Studien 96.1 (2005).

baptism by total immersion in the logic of the categorical imperative itself.[32] Kant describes the conversion as one in which, "by a single and unalterable decision a human being reverses the supreme ground of his maxims by which he was an evil human being," replacing the selfish orientation of the old man with the rational universalizability that characterizes the Kantian new man. "So long as the foundation of the maxims of the human being remains impure," Kant explains, the change "cannot be effected through gradual *reform* but must rather be effected through a *revolution* in the disposition of the human being (a transition to the maxim of holiness of disposition)" (Kant R 6:47–48). This secularized picture of justice would scarcely be recognizable to the more traditional Christian understandings of Anselm—a fact that throws into sharp relief the degree to which God as traditionally conceived does not figure comfortably in Kant's theology. Instead, God's penchant for both rewarding and punishing poses a constant threat to turn all human action into a quest for rewards that is in the Kantian scheme the very antithesis of moral behavior.

In Kant's view, even the possibility of divine mercy is deeply problematic. Given Kant's larger ethical framework, mercy is only justifiable if sinners are somehow able to make satisfaction for themselves, even though this seems logically impossible. Kant's three premises—that the standards of human and divine justice must be the same, that God's justice is retributive, and that moral debts are not transmissible—leads him to the conclusion that more orthodox thinkers had avoided for centuries: that the underlying logic of divine retributivism applies with equal force to the adjudication of human punishment, and that consequently mercy is itself by its very nature a form of injustice.

Conclusion

Anselm's theology held that it would be inconsistent with God's justice if sin were simply forgiven without satisfaction being made. This claim had the potential, if applied to the punishments of human governments, to justify retributivist punishment. The tradition of which Anselm was a part, a tradition stretching back to Augustine and forward to Luther and Calvin, did not in fact make that inference, largely because those in the tradition accepted the Augustinian asymmetry framework. Because God will one day punish according to desert with perfect wisdom and justice, Anselm's successors believed, human beings with their more limited wisdom need not take this retributive responsibility upon themselves. Even within Anselm's framework of necessary divine retributivism, temporal rulers could justify showing mercy to earthly offenders on the grounds that the Trinitarian God had given the incarnate Son as an act of mercy that still satisfied the demands of justice—and had indeed called human beings to do the same. Thinkers with less confidence in these mysteries would be less inclined to maintain the asymmetry.

We have described how in the seventeenth and eighteenth centuries a secularized and politicized form of Anselm's theological framework emerged. The seventeenth century natural law

[32] Indeed, for Kant this "conversion" describes a general state of mind necessary for any moral action, rather than a specific biographical moment of "change" in the more familiar sense. See further Allen W. Wood, *Kant's Moral Religion* (Ithaca, N.Y.: Cornell University Press, 1970), 226–231.

theorists, led by Grotius, denied the central premise of the Augustinian asymmetry, namely that human and divine justice are qualitatively different. They did this not to make human punishment seem more retributive but instead to make divine punishment seem less so. Once this move was made, however, it is unsurprising that in the eighteenth century those more sympathetic to retributive punishment perceived the possibilities inherent in making the opposite move, changing their account of human punishment to correspond with the more traditionally retributive view of divine justice. This movement culminated in a Kantian theory that closely followed Anselm's account of necessity of justice, while simultaneously jettisoning the divine mysteries which had underpinned the Augustinian asymmetry. Modern retributivism thus has its roots in Kant's secularization of these dimensions of Anselm's theology.